Research Note
RM-RN-536
September 1995
USDA Forest Service
Rocky Mountain Forest and
Range Experiment Station

Stratification and Plot Selection Rules: Misuses and Consequences

Hans T. Schreuder[1] and Jim Alegria[2]

If the probabilities of selection for units in a sample of a population are unknown, one should not use the data to draw statistical inferences about the population. To illustrate, we examine questionable practices in forest inventory and derive equations showing the bias that these practices can generate.

Keywords: Unknown probabilities, bias, inference

INTRODUCTION

It is well known in forest inventory practice that stratification of the population of interest is an effective tool both in reducing the variance in estimation and in providing information about the strata themselves (Schreuder et al. 1993). However, its use can be abused. Suppose, for example, the strata are changed after additional information becomes available. This can lead to serious bias in the estimation procedure if the wrong probabilities of selection are used.

Similarly, other sample selection rules may be applied that appear to be practical and useful but may result in unequal probabilities of selection. Often, if this is not taken into account in estimation, serious bias will result. In some cases, the sampling rules used in the past are unclear so the true probabilities of selecting the sampling units are unknown. Using these data because they are the best available can result in unrealistic results.

The purpose of this article is to document the danger in inappropriate estimation if probabilities of selection are assumed equal when they are not.

REVIEW OF LITERATURE

Both purposive sampling and probabilistic sampling have their uses. In purposive sampling particular sample units are selected because the investigator thinks he or she knows what sample best represents the population. This means that the other units in the population have zero probability of being included in the sample. Such a sample can be useful when a quick decision needs to be made (Schreuder and Wood 1986, Schreuder et al. 1993, Ch. 6, and especially the example in Schreuder 1995 based on an example by D. Basu). In probabilistic sampling all units have a positive probability of sampling and these probabilities should be used in estimation for scientific validity.

When probabilities of selection of sample units are known and used correctly in estimation, totals for the variables of interest measured on those units can be

[1] *Math Statistician, Rocky Mountain Range and Experiment Station, Fort Collins, CO*

[2] *Biometrician, Bureau of Land Management, Portland, OR*

estimated unbiasedly. Problems arise when (a) these probabilities are not known because sample selection is changed in an unquantifiable way in the field, (b) are assumed known when in fact the process of how units were selected is now unknown, or (c) assumptions in the selection process are wrong.

Williams et al. (1995) documented the potential estimation bias in moving a cluster of subplots into the condition class of the central subplot. This procedure was used by some Forest Inventory and Analysis (FIA) units of the USDA Forest Service but is now unacceptable. The contention that this bias is not important because the estimates have never been questioned is untenable. We live in a contentious age and scientific credibility needs to be maintained. Finding significant bias in one study undermines the results of other similar studies showing little or no bias.

A method used in timber cruising about 15 years ago by some National Forests personnel was to change the basal area factor in variable radius plot (VRP) sampling if insufficient sample trees were selected at a sample point. The idea was to ensure an adequate count of, say, 4-8 trees per point. This approach was used in several other cases in the western United States because some biometricians found in empirical studies that the resulting bias was negligible. However, in one case in California the procedure was found to incur appreciable bias in estimating wood volume because the actual probabilities of selection were modified in an unquantifiable way and differed from those assumed in estimation (Wensel et al. 1980, Schreuder et al. 1981).

AN EXAMPLE

Emphasis in the past, when sampling public lands in the West, was on the timber resources. The usual scenario was for the commercial forest land base to be divided into several age classes (strata). Each stratum was composed of homogeneous stands of areas ranging from a few to several hundred acres and each stand was uniquely defined and had a known acreage attached to it. All strata were sampled identically. An initial set of plots was used to estimate the variability within each stratum and to estimate the total number of plots necessary to meet a predetermined precision level. The plots were distributed via an optimal allocation formula producing unequal plot selection probabilities within and between the strata.

Ten years later, the population was re-stratified from three to four strata for several reasons. Access to better aerial photography and escalating demands for finer class separation led to more strata. Some stands were harvested, moving them from the oldest to the youngest class, and natural growth of the forest shifted stands to another stratum. The size of the population also changed between the two time periods, reflecting additional acreage due to advances in regeneration techniques and deletions of acreage for a variety of non-timber related reasons. Most of the original plots were remeasured and additional plots were optimally allocated to the four strata based on the assumption that the original plots had an equal probability of selection in the new strata. The estimated volumes and variances were computed assuming a stratified random sample based on the new stratification. This resulted in a situation where the probabilities of selecting the plots in the sample were unequal ***and*** unknown.

Formulas for bias incurred by assuming equal probabilities of selection are derived in the next section. Readers who are not interested in the statistical details can skip this section.

STATISTICAL DEVELOPMENT

Assume three (k_1) age-class strata within which n(1) sample plots were selected and distributed by optimal allocation. Within a stratum, units were selected with equal probability but, to consider a more general case, the probability of selecting unit i in stratum k_1 was $\Pi_{k\ell i}$.

A proportion of the n(1) sample units are remeasured at time two and additional sampling units are also selected, to give n sample units. These n units (plots) are assigned to one of four new (k_2) strata. For simplicity, we assume that the actual k_2 strata sizes are known without error, although this is not always true: Errors would make it more difficult to derive good estimates of estimation bias.

The probabilities of reselecting some or all of the original n_{k1} sample units is clearly Π_{k1} times some factor that reflects both the percentage of units remeasured and which of the k_1 initial strata they fell in. New units would have probabilities $\Pi_{k2\ell j}$ of selection for stratum ℓ (ℓ=1, ..., k_2) (provided the new units are selected independently from the old ones). Because we allow different probabilities of selection

within the new k_2 strata, we denote these probabilities by $\Pi_{k2\ell j}$ (i=1, ..., n) for all sample units, either new or remeasured.

Let n_ℓ units fall in stratum ℓ (ℓ=1, …, k_2), $N = \sum_{\ell=1}^{k_2} n_\ell$, and assume we want to estimate Y_ℓ = total wood volume in stratum ℓ. Given this situation, our unbiased estimators of volume at the current time are:

$$\hat{Y}_\ell^* = \sum_{j=1}^{n_i} y_{\ell i} / \Pi_{k2\ell j} \qquad [1]$$

where $Y_{\ell i}$= value of variable y in stratum ℓ, observation j, and

$$\hat{Y}^* = \sum_{\ell=1}^{k_2} \hat{Y}_\ell. \qquad [2]$$

But the probabilities may no longer be known, especially for units established at time one due to poor record keeping. In these situations, it is often assumed that the units within each stratum have the same probability of selection. Under this scenario, estimators by stratum and overall are:

$$\hat{Y}_\ell' = \left[\sum_{i=1}^{n_i} y_{\ell i} / n_\ell\right] N_\ell = \sum_{i=1}^{N_i}\left[\frac{y_{\ell i}}{n_\ell} * \delta_{\ell i}\right] N_\ell \qquad [3]$$

where $\delta_{\ell i}$ = 1 with probability $\Pi_{k2\ell i}$
= 0 with probability 1 - $\Pi_{k2\ell i}$,

$$\sum_{j=1}^{N_1} \delta_{\ell i} = n_\ell$$

and

$$\hat{Y}' = \sum_{\ell=1}^{k_2} \hat{Y}_\ell. \qquad [4]$$

These estimators can be quite seriously biased due to the earlier stated emphasis on timber volume estimation and optimal allocation to "preferred" strata.

The expected value (E) of $\hat{Y}_\ell'$ can be shown to be

$$E(\hat{Y}_\ell') = \sum_{i=1}^{N_i} \frac{y_{\ell i}}{n_\ell} \Pi_{k2\ell i} N_\ell \qquad [5]$$

which is unbiased only if $\Pi_{k2\ell i} = n_\ell / N_\ell$. The estimated bias of $\hat{Y}_\ell'$ then is

$$\begin{aligned} \hat{B}_\ell &= E(\hat{Y}_\ell) - Y_\ell \qquad [6] \\ &= \sum_{i=1}^{N_i} \frac{y_{\ell i}}{n_\ell} \Pi_{k2\ell i} N_\ell - \sum_{i=1}^{n_i} y_{\ell i} \\ &= \sum_{i=1}^{N_i} y_{\ell i}\left(\frac{\Pi_{k2\ell i}}{n_\ell} N_\ell - 1\right) \end{aligned}$$

To illustrate the magnitude of bias, assume we use

$$\hat{Y}' = \sum_{i=1}^{k_2} \hat{Y}_\ell'$$

with

$$\hat{Y}_\ell' = \left[\sum_{i=1}^{n_\ell} y_{\ell i}/n_\ell\right] N_\ell,$$

when in fact we should have used

$$\hat{Y}^* = \sum_{\ell=1}^{k_2} \hat{Y}_\ell^*$$

with

$$Y_\ell = \sum_{j=1} y_{\ell i} / \Pi_{k2\ell j}.$$

For simplicity, we use $\Pi_{2\ell j}$ instead of $\Pi_{2k\ell j}$.

Then the bias (B) of estimator $\hat{Y}$ in estimating parameter Y is (assuming k_2=1)

$$\begin{aligned} B &= E(\hat{Y}') - E(\hat{Y}^*) = \sum_{i=1}^{N} y_i\left(\tfrac{N}{n}\Pi_{2i} - 1\right) \qquad [7] \\ &= \frac{N}{n} \sum_{i=1}^{N} y_i \Pi_{2i} - Y \end{aligned}$$

If p percent of the plots were selected with m times the probability of the other (1-p)N plots (m=2, 3, ...), then since $\sum_{i=1}^{N} \Pi_{2i} = n$,

$$\sum_{i=1}^{pN} \Pi_{2i} + \sum_{j=1}^{N-pN} \Pi_{2j} = n.$$

Then if Π_{2i} = m Π_{2j} = Π, (i.e. $\Pi_{2i} = \Pi$, $\Pi_{2j} = \frac{\Pi}{m}$)

$$\Pi = \frac{nm}{N(mp - p + 1)}.$$

So B can be rewritten as

$$B = \frac{N}{n}\left[\sum_{i=1}^{pN} y_i \frac{nm}{N(mp-p+1)} + \sum_{j=1}^{N-pN} y_j \frac{n}{N(mp-p+1)}\right] - Y$$
$$= \sum_{i=1}^{pN} \frac{my_i}{(mp-p+1)} + \sum_{j=1}^{N-pN} \frac{y_j}{(mp-p+1)} - Y \qquad [8]$$
$$= \sum_{i=1}^{pN} \frac{(m-1)y_i}{(mp-p+1)} + \frac{Y}{(mp-p+1)} - Y.$$

The bias can be positive or negative depending on whether the higher probabilities are associated with the larger or smaller y_i values. For example, for y = volume, bias can be expected to be positive if the original designs emphasized larger sample sizes in the strata with larger trees. Similarly, one might expect negative bias for mortality since it is often associated with smaller trees.

The above can be illustrated with a simple application of [8]. Assume p = 1/2, then

$$B = \sum_{i=1}^{N/2} \frac{(m-1)y_i}{\frac{1}{2}(m+1)} + \frac{Y}{\frac{1}{2}(m+1)} - Y. \qquad [9]$$

Now to simplify [9], assume m = 3, then

$$B = \sum_{i=1}^{N/2} y_i + \frac{Y}{2} - Y = \sum_{i=1}^{N/2} y_i - Y/2. \qquad [10]$$

If the larger y_i values are selected with probability Π, and if we substitute this knowledge into [10], we get

$$B = \sum_{i=1}^{N/2} y_i - \frac{Y}{2} < \frac{Y}{2} - \frac{Y}{2} = 0$$

and if the smaller y_i-values are selected with probabilities Π, then

$$B = \sum_{i=1}^{N/2} y_i - \frac{Y}{2} > \frac{Y}{2} - \frac{Y}{2} = 0.$$

PRACTICAL IMPLICATIONS

To provide some indication of how serious the bias B can be, assume that the largest units are selected with three times the probability of smaller units (m=3) and that the average size of the N/2 largest units is twice that of the N/2 smallest units. Both assumptions are not unreasonable for traditional timber-oriented surveys. Substituting bias information into [10] where the N units are sorted by y_i values in decreasing order, and noting that the overall total is $Y = N\left[\frac{\bar{Y}_1+\bar{Y}_2}{2}\right]$, where $\bar{Y}_1$ is the mean for the N/2 largest and $\bar{Y}_2$ the mean for the N/2 smallest units, we obtain

$$B = \frac{N}{2}\bar{Y}_1 - \frac{N\bar{Y}_1 + N\bar{Y}_2}{4}$$
$$= -\frac{N}{4}\bar{Y}_2.$$

Since $\bar{Y}_1 = 2\bar{Y}_2$, $Y = \frac{3N\bar{Y}_2}{2}$ and if we express B as a percent of Y, we get $B(\%) = (-\frac{N}{4}\bar{Y}_2 / Y)*100\% \doteq 17\%$.

Similar serious biases can occur if the smaller y_i values are selected with higher probability.

SUMMARY

We have shown for a simplified but realistic situation that bias in estimation can be serious if the probabilities of selecting the sampling units are ignored. Our recommendation is not to use data sets for inference when these probabilities are unknown.

REFERENCES

Schreuder, H. T. 1995. Simplicity versus efficiency in sampling designs and estimation. Environmental Monitoring and Assessment 33:237-245.

Schreuder, H. T., Gregoire, T. G., and Wood, G. B. 1993. Sampling methods for multiresource forest inventory. New York: John Wiley & Sons, 446 pp.

Schreuder, H. T.; Schreiner, D. A.; and Max, T. E. 1981. Ensuring an adequate sample at each location in point sampling. Forest Science 27: 567-578.

Schreuder, H. T. and Wood, G. B. 1986. The choice between design-dependent and model-dependent sampling. Canadian Journal of Forest Research 16: 260-265.

Wensel, L.; Levitan, J.; and Barber, K. 1980. Sample selection of basal area factor in point sampling. Journal of Forestry 78: 83-84.

Williams, M. S.; Schreuder, H. T.; and Reich, R. 1995. Bias due to rotation in forest survey. USDA Forest Service RM Station Paper.